Medical Image Analysis 62 (2020) 101688

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI

Dennis Bontempi[a], Sergio Benini[a], Alberto Signoroni[a], Michele Svanera[b,1,*], Lars Muckli[b,1]

[a] Department of Information Engineering, University of Brescia, Brescia, Italy
[b] Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom

ARTICLE INFO

Article history:
Received 30 September 2019
Revised 7 March 2020
Accepted 12 March 2020
Available online 24 March 2020

Keywords:
Brain MRI segmentation
Convolutional neural networks
Weakly supervised learning
3D Image analysis

ABSTRACT

Many functional and structural neuroimaging studies call for accurate morphometric segmentation of different brain structures starting from image intensity values of MRI scans. Current automatic (multi-) atlas-based segmentation strategies often lack accuracy on difficult-to-segment brain structures and, since these methods rely on atlas-to-scan alignment, they may take long processing times. Alternatively, recent methods deploying solutions based on Convolutional Neural Networks (CNNs) are enabling the direct analysis of out-of-the-scanner data. However, current CNN-based solutions partition the test volume into 2D or 3D patches, which are processed independently. This process entails a loss of global contextual information, thereby negatively impacting the segmentation accuracy. In this work, we design and test an optimised end-to-end CNN architecture that makes the exploitation of global spatial information computationally tractable, allowing to process a whole MRI volume at once. We adopt a weakly supervised learning strategy by exploiting a large dataset composed of 947 out-of-the-scanner (3 Tesla T1-weighted 1mm isotropic MP-RAGE 3D sequences) MR Images. The resulting model is able to produce accurate multi-structure segmentation results in only a few seconds. Different quantitative measures demonstrate an improved accuracy of our solution when compared to state-of-the-art techniques. Moreover, through a randomised survey involving expert neuroscientists, we show that subjective judgements favour our solution with respect to widely adopted atlas-based software.

1. Introduction

The segmentation of various brain structures from MRI scans is an essential process in several non-clinical and clinical analyses, such as the comparison at various stages of normal brain, or disease development of neurodegenerative processes, neurological diseases, and psychiatric disorders. The morphometric approach is especially helpful in pathological situations for confirming the diagnosis, defining the prognosis, and selecting the best treatment. Moreover, brain structure segmentation is an early step in functional MRI (fMRI) study pipelines, as neuroscientists need to isolate specific brain structures before analysing the spatiotemporal patterns of activity within them.

Manual segmentation, although considered to be the gold standard in terms of accuracy, is time consuming (Zhan et al., 2018). Therefore, neuroscience studies began to exploit computer vision to process data from increasingly performing MRI scanners and ease the interpretation of brain data, intrinsically characterised by a strong inter-subject variability. Different fully automated pipelines have been developed in recent years (Despotović et al., 2015), moving from techniques based only on image features to ones that make also use of a-priori statistical knowledge about the neuroanatomy. The vast majority of the available tools apply a (multi-) atlas-based segmentation strategy (Cabezas et al., 2011), in which the segmentation of the target volume is inferred from one or several templates built from manual annotations. In order to make this inference phase possible, a time consuming and computationally intensive (FreeSurfer, 2008) non-rigid subject-to-atlas alignment is necessary. Due to the aforementioned high inter-subject brain variability, such registration procedures often introduce errors that yield a decrease in segmentation accuracy on

* Corresponding author.
E-mail address: michele.svanera@glasgow.ac.uk (M. Svanera).
[1] Shared authorship.

https://doi.org/10.1016/j.media.2020.101688

brain structure or tissue boundaries (Klein et al., 2017; Lerch et al., 2017).

In recent years, Deep Learning (DL) techniques have emerged as one of the most powerful ways to combine statistical modelling of the data with pattern recognition for decision making and classification (Voulodimos et al., 2018), and their development is impacting various medical imaging domains (Hamidinekoo et al., 2018; Litjens et al., 2017). Provided that they are trained on a sufficient amount of data embodying the observable variability, DL models are able to generalise well to previously unseen data. Furthermore, they can work directly with out-of-the-scanner images, removing the need for the expensive scan-to-atlas alignment phase. Numerous DL-based algorithms proposed for brain MRI segmentation match or even improve the accuracy of atlas-based segmentation tools (Akkus et al., 2017; Rajchl et al., 2018; Roy et al., 2019; Wachinger et al., 2018). Due to the scarcity of training data and to hardware limitations, approaching this task using DL commonly requires the volume to be processed considering 2D (Roy et al., 2019) or 3D-patches (Fedorov et al., 2017; Rajchl et al., 2018; Dolz et al., 2019; Wachinger et al., 2018; Li et al., 2017) at a time. Although this method simplifies the process from a technical point of view, it introduces significant limitations in the analysis: since each 2D or 3D patch is segmented independently from the others, these models mostly exploit local spatial information - ignoring "global" cues, such as the absolute and relative positions of different brain structures - which makes them sub-optimal. Different works have considered the potential improvements of removing said volume partitioning (McClure et al., 2018; Wachinger et al., 2018). Solutions that exploit such fully-volumetric approach have already been applied to prostate (Milletari et al., 2016), heart atrium (Savioli et al., 2019), and proximal femur MRI segmentation (Deniz et al., 2018), but not yet to brain MRI segmentation - where this strategy could prove particularly useful given the complex geometry and the variety of structures characterising the brain anatomy. Here, we discuss how both hardware limitations and the scarcity of hand-labelled ground truth (GT) data can be overcome. First, we tackle the former by customising and simplifying the model architecture. Second, the latter is coped with by training our model on segmentation masks obtained exploiting atlas-based techniques, in what can be considered a weakly supervised fashion - more precisely what (Zhou, 2017) and (Li et al., 2019) describe as "inaccurate supervision". Hence, even though CEREBRUM is trained exploiting labelling which is not exempt from errors, we demonstrate that the statistical reliability of atlas-based segmentation is enough to guarantee good generalisation capability of the DL models trained on such imperfect ground truth.

2. Existing methods for whole brain MRI segmentation and how to advance them

2.1. Atlas-based methods

In the last twenty years, several atlas-based segmentation methods have been developed. However, only a few of them are completely automatic, and thus pertinent to our discussion: FreeSurfer, FSL's FAST and FMRIB, and fMRIprep. FreeSurfer (Fischl, 2012) is an open-source software package that contains a completely automated pipeline for tissue and sub-cortical brain structure segmentation. FSL's FAST (FMRIB's Automated Segmentation Tool, Zhang et al., 2001) and FIRST (FMRIB's Integrated Registration and Segmentation Tool, Patenaude et al., 2011) are part of the Oxford's open-source library of analysis tools for MRI and fMRI data. FAST segments different tissue types in already skull-stripped brain scans, while FIRST deals with the segmentation of sub-cortical brain structures. fMRIprep (Esteban et al., 2019) is a recently published preprocessing software for MRI scans that combines tools from widely used open-source neuroimaging packages (e.g., the above mentioned FSL and FreeSurfer). It implements a brain tissues segmentation pipeline, providing the user with both soft (i.e., probability maps) and hard segmentation.

These methods are widely used in neuroscience, since they produce consistent results with little human intervention. Nevertheless, they are all atlas-based and not learning-based - hence, the only way to improve their accuracy is to manually produce new atlases. Furthermore, since they implement a long processing pipeline together with the atlas-based labelling strategy, the segmentation operation is time consuming (FreeSurfer, 2008). Limitations of these approaches, such as the lack of accuracy on various brain structure boundaries, have been documented (Ellingsen et al., 2016; Wenger et al., 2014; Weier et al., 2012; Cabezas et al., 2011).

2.2. Deep learning methods

Many of the state-of-the-art methods based on deep learning exploit multi-modal MRI data (Çiçek et al., 2016; Chen et al., 2018; Dolz et al., 2019; Andermatt et al., 2016). Yet, in real-case scenarios and due to time constraints, the acquisition of different MRI sequences for anatomical analysis is rarely done: in most studies a single sequence is used - with $T1_w$ being the most popular protocol. Various alternatives have been proposed to obtain whole brain segmentation from $T1_w$ only. QuickNAT (Roy et al., 2019) leverages a 2D based approach to efficiently segment brain MRI, exploiting a paradigm that aggregates the predictions of three different encoder-decoder models by averaging the probability maps - each model trained to segment a single slice at a time along one of the three principal axes (longitudinal, sagittal, and coronal). MeshNet (Fedorov et al., 2017; McClure et al., 2018) is a feedforward CNN based on 3D dilated convolutions, whose structure guarantees good results while keeping the number of parameters low. NeuroNet (Rajchl et al., 2018) is an encoder-multi-decoder CNN, trained to replicate segmentation results obtained with multiple state-of-the-art neuroimaging tools. DeepNAT (Wachinger et al., 2018) is composed of a cascade of two CNNs. It breaks the segmentation task into two hierarchical operations - the foreground-background separation, and the labelling of each voxel as belonging to the foreground - implemented by the first and the second network, respectively. Finally, the solution presented in Li et al. (2017) makes use of various refinements, such as residual connections and dilated convolution, to favour the learning of 3D representation and increase the compactness of the proposed model. Such modifications are furthermore at the centre of the extensive analysis conducted by the authors in an effort to explain how the former impact the model performance.

However, a common trait of these methods is that they do not fully exploit the 3D spatial nature of MRI data. Although QuickNAT tries to integrate spatial information by averaging the probability maps computed with respect to different views, it is slice-based. DeepNAT exploits an intrinsic parameterisation of the brain (through the Laplace-Beltrami operator) trying to introduce some spatial context, but as with MeshNet it is trained on small non-overlapping 3D-patches. Finally, NeuroNet is trained on random crops of the MR volume, and so is the high-resolution compact CNN presented in Li et al. (2017).

2.3. Aims and contributions

Aiming to exploit both local and global spatial information contained in MRI data, we introduce CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly supervised sEgmentation of BRain strUctures from out-of-the-scanner

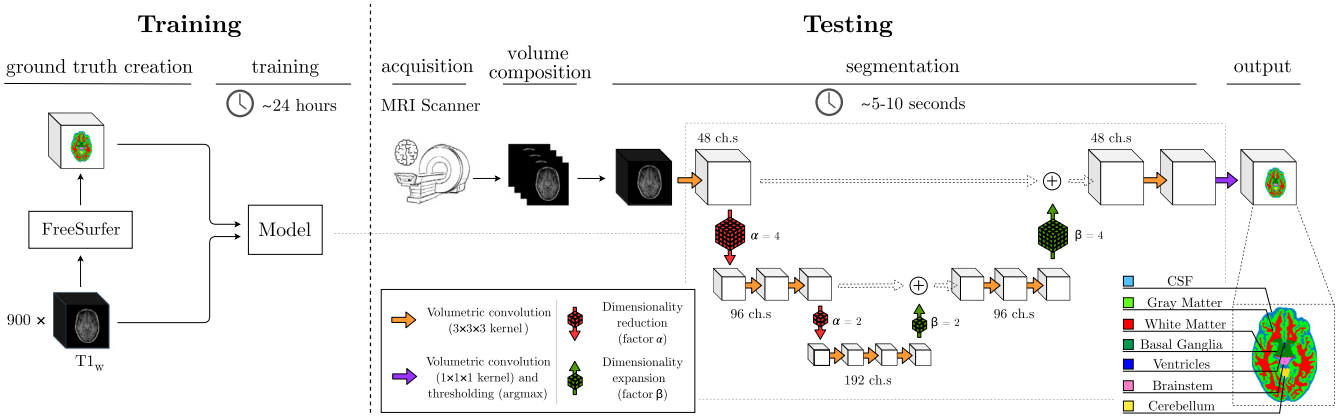


Fig. 1. Overview of the proposed segmentation method. The model is trained on 900 $T1_w$ volumes and the associated relabelled FreeSurfer segmentation, while testing is performed by feeding NIfTI data to the model.

MRI. To the best of our knowledge, CEREBRUM is the first DL model designed to tackle the brain MRI segmentation task in such a fully-volumetric fashion. This is accomplished exploiting an end-to-end encoding-decoding structure, where only convolutional blocks are used. This delivers a whole brain MRI segmentation in just ∼5–10 s on a desktop GPU. The model architecture and the proposed learning framework are shown in Fig. 1.

Since in most real case scenarios, to save scanner time, only single-modal MR images are collected, we develop and test our method on a large set of data (composed by 947 MRI scans) acquired using a T1-weighted ($T1_w$) 1*mm* isotropic MPRAGE protocol. Neither registration nor filtering is applied to these data, so that CEREBRUM learns to segment out-of-the-scanner volumes. Focusing on the requirements of a real case scenario (fMRI studies), we train the model to segment the classes of interest in the MICCAI challenge (Mendrik et al., 2015) i.e., gray matter (GM), white matter (WM), cerebrospinal fluid (CSF), ventricles, cerebellum, brainstem, and basal ganglia. Since manually annotating such a large body of data would require a prohibitive amount of human hours, we train our model on automatic segmentations obtained by FreeSurfer (Fischl, 2012) - relabelled to obtain the aforementioned set of seven classes.

We compare the proposed method with other CNN-based solutions: the well-known 2D-patch-based U-Net (Ronneberger et al., 2015), its 3D variant (Çiçek et al., 2016), and the state-of-the-art architecture QuickNAT (Roy et al., 2019) - which leverages the aggregation of three slightly modified U-Net architectures (trained on coronal, sagittal, and axial MRI slices, respectively). To ensure a fair comparison, we train these models by conducting an extensive hyperparameter selection process. Results are quantitatively evaluated exploiting the same metrics used in the MICCAI MR Brain Segmentation challenge, i.e., the Dice Similarity Coefficient, the 95th Hausdorff Distance, and the Volumetric Similarity Coefficient (Taha and Hanbury, 2015), utilising FreeSurfer as GT reference. In addition, to assess the generalisation capability of the proposed model, we compare the obtained results against the FreeSurfer segmentation we used for training. To do so, we design a survey in which five expert neuroscientists (with more than five years of experience in MRI analysis) are asked to choose the most accurate segmentation between the two aforementioned ones. This qualitative test covers different areas of interest in neuroimaging studies, i.e., the early visual cortex (EVC), the high-level visual areas (HVC), the motor cortex (MCX), the cerebellum (CER), the hippocampus (HIP), the early auditory cortex (EAC), the brainstem (BST) and the basal ganglia (BGA).

All the code necessary to train CEREBRUM and run the survey is available at the project's GitHub page.[2]

3. Material and methods

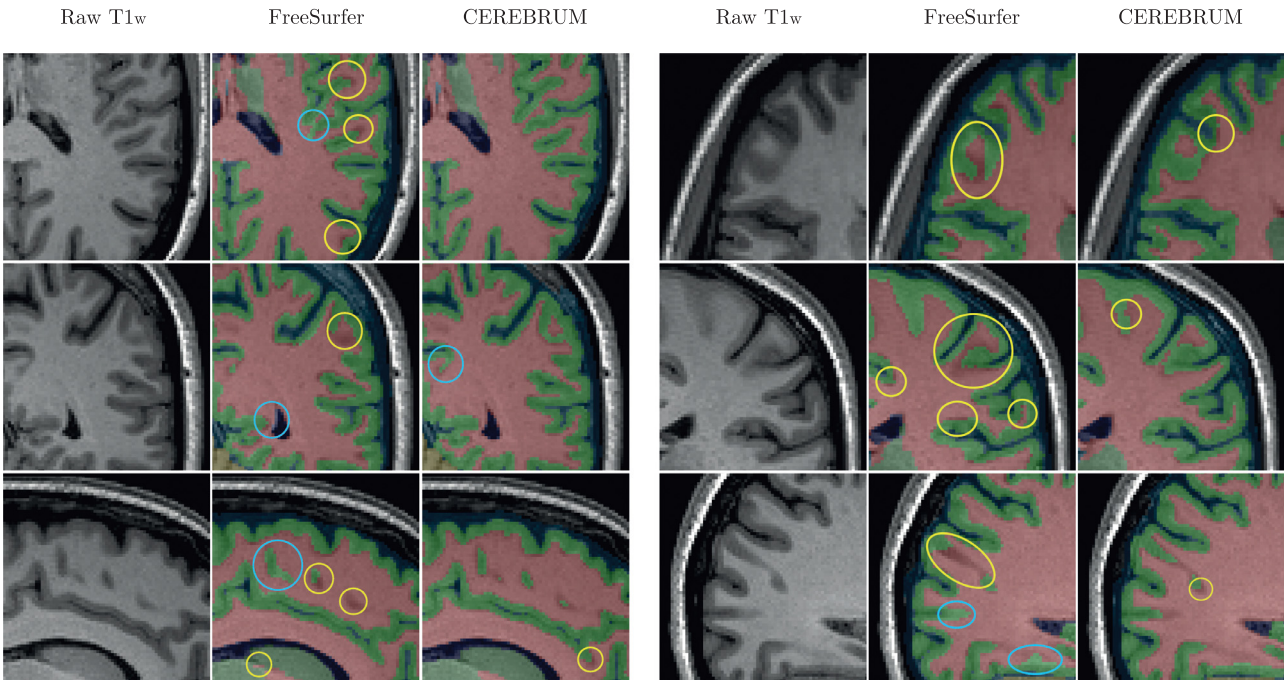
3.1. Data

To speed up research and promote reproducibility, numerous large-scale neuroimaging experiments make the collected data available to all researchers (Marcus et al., 2007; Van Essen et al., 2013; Oxtoby et al., 2019; Miller et al., 2016; Bellec et al., 2017). However, none of these studies provide manual annotations, as carrying out the operation on such large databases would prove exceptionally time-consuming.

For this reason, most of the studies investigating the application of DL architectures for brain MRI segmentation make use of automatically produced GT for training purposes (Roy et al., 2019; McClure et al., 2018; Fedorov et al., 2017; Rajchl et al., 2018) - with some of them reporting the latter can be exploited to train models that perform the same (Rajchl et al., 2018), or even better (Roy et al., 2019), than the automated pipeline itself. Motivated by this rationale, we train and test the proposed model using both a large collection of out-of-the-scanner MR images and the results of the FreeSurfer (Fischl, 2012) cortical reconstruction process, `recon-all`, as reference GT. As anticipated in Section 1, we relabel this result preserving seven among the most important classes of interest in most of fMRI studies (see Section 2.3 and Fig. 1).

The database, collected from the Centre for Cognitive Neuroimaging (the University of Glasgow) in more than 10 years of machine activity, consists of 947 MR images - 900 of which are used for training, 11 for validation, and 36 for testing. All the volumes are out-of-the-scanner, i.e, obtained directly from a set of DICOM images using `dcm2niix` (Li et al., 2016), whose auto-crop option is exploited to make sizes consistent across all the dataset (i.e., $192 \times 256 \times 170$ for sagittal, coronal, and longitudinal axis, respectively) without any other pre-processing of the data. Given the number of available scans for training, and since no registration is performed, the variability in shape, rotation, position, and anatomical size is such that no data augmentation is needed to avoid the risk of overfitting. The first two columns of Fig. 2(a) and (b) show detailed views from some selected slices of the out-of-the-scanner $T1_w$ and the corresponding relabelled FreeSurfer segmentation, respectively. The main characteristics of the dataset are summarised in Table 1. As the data have been collected under different ethics

[2] https://github.com/denbonte/CER3BRUM.



(a) Subject 1 (b) Subject 4

Fig. 2. Out-of-the-scanner (contrast enhanced) $T1_w$ scan (left), FreeSurfer segmentation (middle), and the result produced by our model (right). Fig. (a) depicts slices of test Subject 1, while (b) slices of test Subject 4 (sagittal, coronal, and longitudinal view, respectively). Cases of white matter over-segmentation are highlighted by yellow circles, while cases of white matter under-segmentation are highlighted by turquoise circles (best viewed in electronic format).

Table 1
Datasets details. MR Images acquired at the Centre for Cognitive Neuroimaging (University of Glasgow, UK).

Parameter	Value
Sequence used	$T1_w$ MPRAGE
Field strenght	3 Tesla
Voxel size	1mm-isotropic
Original volume sizes	$192 \times 256 \times 256$
Training volume sizes	$192 \times 256 \times 170$[b]
Training	900 volumes
Validation	11 volumes
Testing	36 volumes[a]

[a] 7 of which are publicly available.
[b] out-of-the-scanner data, neck cropping only.

applications, we are not able to make the whole database publicly available. However, 7 out of 36 volumes used for testing are collected under the approval of the local ethics committee of the College of Science & Engineering (ethics #300170016) and shared online after anonymisation,[3] for comparison and research purposes, along with the segmentation masks resulting from CEREBRUM and FreeSurfer (See Fig. 2 and Section 4.2).

3.2. Proposed model

To make the complexity of managing our $192 \times 256 \times 170$ voxels data tractable, we carefully optimise the model architecture so as to implicitly deal with GPU memory constraints. Furthermore we exploit, for training purposes, a machine equipped with 4 GeForce® GTX 1080 Ti - distributing different parts of the model on different GPUs.

Inspired by Ronneberger et al. (2015) and Çiçek et al. (2016), we propose a deep encoder-decoder model with six 3D convolutional blocks, which are arranged in increasing number on three layers. Since a whole volume is considered as an input, the feature maps extracted by such convolutional blocks are not limited to patches but span across the entire volume. As each block captures the content of the whole brain MRI, this enables the learning of both local and global spatial features by leveraging the spatial context which is propagated to each subsequent block. The capability of CEREBRUM to learn both local and global features is coherent with the last layer units of the model having a 100x100x100 theoretical receptive field. A table reporting the complete calculation of such parameter for each convolutional block of CEREBRUM can be found in the Supplementary Material. In order to better exploit the fine details found in 3T brain MRI data, kernels of size $3 \times 3 \times 3$ are used as feature extractors. Instead of max-pooling, convolutions with stride are used as a dimensionality reduction method, thus allowing the network to learn the optimal down-sampling strategy starting from the extracted features. Exploiting such operations, and to force the learning of more abstract (spatial) features, a factor 1: 64 dimensionality reduction is implemented after the first layer. Finally, skip connections are used along with tensorial sum (instead of concatenation, Quan et al., 2016) to improve the quality of the segmented volume while greatly limiting the number of parameters to $\sim 5M$, far less with respect to state-of-the-art models which are structured in a similar fashion.

We train the model by optimising the categorical cross-entropy function. Convergence is achieved after roughly 24 hours of training (40 epochs), using Adam (Kingma and Ba, 2014) with a learning rate of $42 \cdot 10^{-5}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Furthermore, we set the batch size to 1 and thus do not implement batch normalisation (Ioffe and Szegedy, 2015).

4. Results

The results we present in this section aim to confirm the hypothesis that avoiding the partitioning of MRI data enables the

[3] https://openneuro.org/datasets/ds002207/versions/1.0.0.

CEREBRUM to learn global spatial features useful for improving segmentation. At first, in Section 4.1, we provide numerical comparison with other state-of-the-art CNN architectures (U-Net, Ronneberger et al., 2015; 3D U-Net, Çiçek et al., 2016; QuickNAT, Roy et al., 2019). Then, in Section 4.2, we conduct a survey involving expert neuroscientists to subjectively assess the CEREBRUM segmentation accuracy. Finally, we further verify the validity of our assumptions by inspecting the soft-segmentation maps produced by the models in Section 4.3, and we demonstrate the suitability of our dataset by analysing the impact of the training set size on CEREBRUM performance in Section 4.4.

4.1. Numerical comparison

We numerically assess the performance of the models, using FreeSurfer segmentation as a reference, exploiting the metrics utilised in the MICCAI MRBrainS18 challenge (among the most employed in the literature, Taha and Hanbury, 2015). Dice (similarity) Coefficient (DC) is a measure of overlap, and a common metric in segmentation tasks. The Hausdorff Distance, a dissimilarity measure, is useful to gain some insight on contours segmentation. Since HD is generally sensitive to outliers, a modified version (95^{th} percentile, HD95) is generally used when dealing with medical image segmentation evaluation (Huttenlocher et al., 1993). Finally, the Volumetric Similarity (VS), as in Crdenes et al. (2009), evaluates the similarity between two volumes.

CEREBRUM is compared against state-of-the-art encoder-decoder architectures: the well-known-2D-patch based U-Net (Ronneberger et al., 2015, trained on the three principal views, i.e., longitudinal, sagittal, and coronal), the 3D-patch based U-Net 3D (Çiçek et al., 2016 - with 3D patches sized $64 \times 64 \times 64$, as in Çiçek et al., 2016; Fedorov et al., 2017; Pawlowski et al., 2017), and the QuickNAT architecture (Roy et al., 2019), which implements view-aggregation starting from 2D-patch based models. We train all the models minimising the same loss for 50 epochs, using the same number of volumes, and similar learning rates (with changes in those regards made to ensure the best possible validation score). Fig. 3 shows class-wise results (DC, HD95, and VS) depicting the average score (computed across all the 36 test volumes) and the standard deviation. We compare 2D-patch-based (longitudinal, sagittal, coronal), QuickNAT, 3D-patch-based, and CEREBRUM (both a max pooling and strided convolutions version). Overall, the latter outperforms all the other CNN-based solutions on every class, despite having far less parameters: when its average score (computed across all the subjects) is comparable with that of other methods (e.g., view-aggregation, GM), it has a smaller variability (suggesting higher reliability). Moreover, we determine the p-values for such scores computing a paired t-test using as a reference the strided-convolutions version of CEREBRUM. In Fig. 3, statistically significant findings ($p < 0.05$) are highlighted with asterisks, whereas the numerical results are reported in the Supplementary Materials.

4.2. Experts' qualitative evaluation

The quantitative assessment presented in Section 4.1, though informative, cannot be considered exhaustive. Indeed, using FreeSurfer as a reference for such evaluation makes the latter a ranking on a relative scale - and if this highlights the value of the fully-volumetric approach, it does not make a direct comparison with the atlas-based method possible. Thus, we need to confirm more systematically what can be inferred, for instance, from Fig. 2 - where far superior qualitative performance of CEREBRUM are clear compared to FreeSurfer, as the former produces more accurate segmentation masks, with far less holes and bridges. This somehow surprising generalisation capability of CEREBRUM over its training reference, if confirmed, would prove the desired "strengthening" effect yielded by the adoption of a weakly supervised learning approach. Moreover, quantitative assessments are often criticised by human experts, such as physicians and neuroscientists, for they do not take into account the severity of each segmentation error (Taha and Hanbury, 2015), which is of critical importance in professional usage scenarios.

For the aforementioned reasons, we design and implement a systematic subjective assessment by means of a PsychoPy (Peirce, 2007) test in which five expert neuroscientists (with more than five years of expertise in MRI analysis) are asked to choose the most accurate segmentation between the one produced by CEREBRUM and the (relabelled) FreeSurfer one. The participants are presented with a coronal, sagittal, or axial slice selected from a test volume, and are allowed both to navigate between four neighbouring slices (two following and two preceding the displayed one) and to change the opacity of the segmentation mask (from 0% to 100%) to better evaluate the latter with respect to the anatomical data. This process is repeated seven times - one for each test subject - per each of the eight brain areas of interest, i.e., early visual cortex (EVC), the high-level visual areas (HVC), the motor cortex (MCX), the cerebellum (CER), the hippocampus (HIP), the early auditory cortex (EAC), the brainstem (BST), and the basal ganglia (BGA). The choice of the slices to present and the order in which the latter are arranged is randomised. Furthermore, the neuroscientists are allowed to skip as many slices as they want if they are unsure about the choice: such cases are reported separately. The survey interface and a run example are provided in the Supplementary Material. From the results shown in Fig. 4 it emerges that, according to expert neuroscientists, CEREBRUM qualitatively outperforms FreeSurfer. This proves the model superior generalisation capability and provides evidence to support the adopted weakly supervised approach. Moveover, such results hint at the possibility to have atlas-based methods and deep learning ones operating together in a synergistic way.

4.3. Probability maps and entropy measures

To further investigate the hypothesis that a fully-volumetric approach is advantageous with respect to other patch-based models, we also conduct an assessment on the predicted probability maps (i.e., soft segmentation). Such evaluation could clearly reveal the ability of the model to make use of spatial cues: for instance, a well-learned model which exploits learned spatial features should predict the presence of cerebellum voxels only in the back of the brain, where the structure is normally located.

Fig. 5(a) and (b) show two selected slices of the soft segmentation (percent probability, displayed in logarithmic scale) resulting from the best 2D-patch-based method (i.e., QuickNAT), the 3D-patch-based method, and CEREBRUM - for the cerebellum and basal ganglia classes, respectively (superimposed to the corresponding $T1_w$ slice). Other classes are omitted for clarity.

The probability maps produced by the 2D and 3D-patch based methods are characterised by the presence of voxels associated with significant probability of belonging to the structure of interest ($p > 0.2$) despite their distance from the latter. This can lead to misclassification errors in the hard segmentation (after the thresholding). In particular, higher uncertainty and spurious activations due to views averaging can be seen in the soft segmentation maps produced by QuickNAT - while blocking artefacts on the patch borders are visible in the case of the 3D-U-Net, even when the latter is trained using overlapping 3D-patches whose predictions are then averaged. The soft segmentation produced by CEREBRUM, on the contrary, is more coherent and closer to the reference in both cases, and does not present the aforementioned errors.

Fig. 3. Dice Coefficient, 95th percentile Hausdorff Distance, and Volumetric Similarity computed using FreeSurfer relabelled segmentation as a reference. The 2D-patch-based (red, green, blue, and grey for longitudinal, sagittal, coronal, and view-aggregation, respectively), the 3D-patch-based (pink), and our model (yellow for max-pooling and orange for strided convolutions) are compared. The height of the bar indicates the mean across all the test subjects, while the error bar represents the standard deviation. The asterisks below the bars highlight statistically significant results ($p < 0.05$), where the p-value is obtained from a paired t-test computed with respect to the strided-convolutions version of CEREBRUM, labelled with "ref." (best viewed in electronic format).

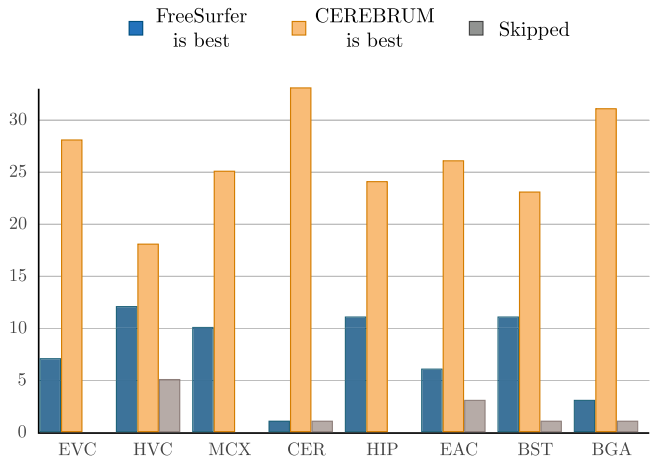


Fig. 4. Outcome of the segmentation accuracy assessment test, conducted by expert neuroscientists, for the following areas: early visual cortex (EVC), the high-level visual areas (HVC), the motor cortex (MCX), the cerebellum (CER), the hippocampus (HIP), the early auditory cortex (EAC), the brainstem (BST), and the basal ganglia (BGA). The bars represent the number of preferences expressed by the experts: CEREBRUM (in orange), FreeSurfer (in blue), or none of the two (in grey). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Beside such qualitative evaluations, we quantitatively compare the sparseness of the predicted probability maps exploiting the average voxel-wise entropy H_V, defined as:

$$H_V(V, \text{DNN}) = \frac{\sum_{v \in V} H_v(v, \text{DNN})}{|V|} \tag{1}$$

where V is an MRI volume, DNN a trained DL model, and v a single voxel. Hence, $|V|$ is the total number of voxels in the volume, and the summation in Eq. (1) is computed for every voxel v in V. The quantity H_v is the voxel-wise entropy, defined starting from the classical definition in Shannon (1948):

$$H_v(v, \text{DNN}) = -\sum_{c \in C} \mathbb{P}(\text{DNN}(v) \in c) \ln\left(\mathbb{P}(\text{DNN}(v) \in c)\right) \tag{2}$$

where C is the set of the segmented classes (in our case $C = \{\text{GM}, \text{GANGL}, \ldots, \text{BRNSTEM}\}$), $\mathbb{P}(\text{DNN}(v) \in c)$ is the probability the model assigns to the event "voxel v belongs to the class c", and $\ln\left(\mathbb{P}(\text{DNN}(v) \in c)\right)$ is the natural logarithm of such quantity. We report the results of such test in Fig. 6(a), "normalising" the quantity H_V by the highest average voxel-wise entropy achievable H_V^{MAX}, i.e., H_V computed for a voxel for which every class is predicted as equiprobable - so that $H_V/H_V^{MAX} \in [0, 1]$ for ease of interpretation.

If entropy evaluates the sparseness of the predicted probability maps in general, cross-entropy is able to assess the uncertainty of a model, once the correct predictions are known. The average voxel-wise cross-entropy CH_V builds upon the idea of the voxel-wise entropy CH_v, defined as:

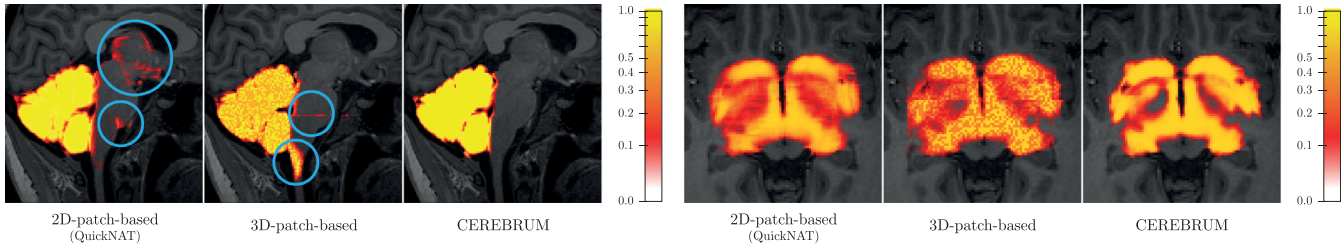
$$CH_v(v, \text{GT}, \text{DNN}) = -\sum_{c \in C} \mathbb{P}(\text{GT}(v) \in c) \ln\left(\mathbb{P}(\text{DNN}(v) \in c)\right) \tag{3}$$

where GT is the ground-truth reference provided by FreeSurfer, C is the set of the segmented classes, $\mathbb{P}(\text{GT}(v) \in c)$ is the probability related to the ground-truth event "voxel v belongs to the class c" (i.e., "1" for the correct class, and "0" otherwise, since FreeSurfer does not provide class probabilities), and $\ln\left(\mathbb{P}(\text{DNN}(v) \in c)\right)$ is the natural logarithm of the probability the DNN model assigns to the event "voxel v belongs to the class c". We report the results of such test in Fig. 6(b), "normalising" the quantity CH_V by the highest average voxel-wise entropy achievable CH_V^{MAX}, i.e., the cross-entropy computed by using a random classifier - so that $CH_V/CH_V^{MAX} \in [0, 1]$.

Both qualitative examples illustrated in Fig. 5, and quantitative evaluations presented in Fig. 6, hint at the superior ability of the proposed model in learning both global and local spatial features. Additional qualitative examples of probability maps, as well as the tables reporting the p-values for the tests depicted in Fig. 6, are provided in the Supplementary Material.

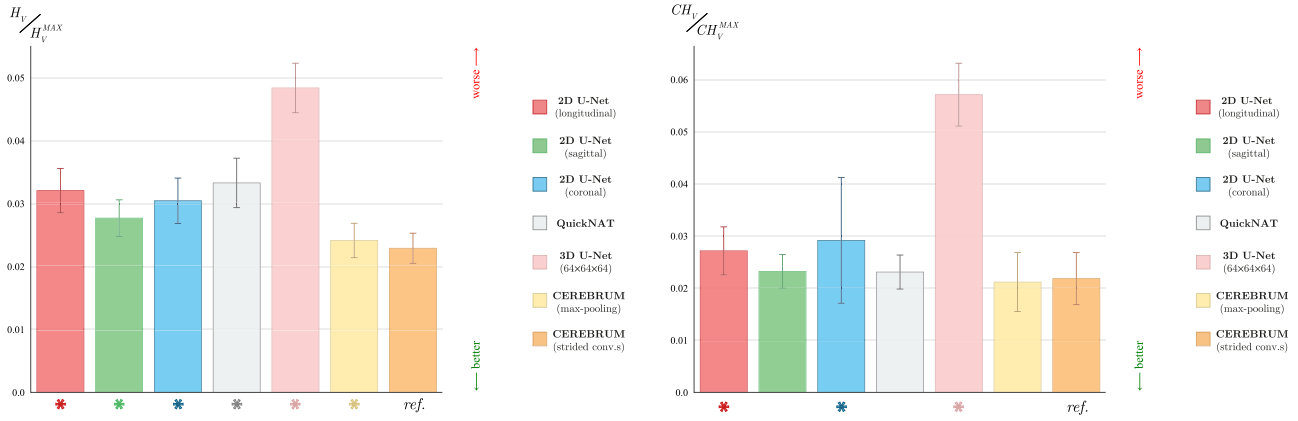
4.4. Number of training samples

One of the possible limitations of approaching the brain MRI segmentation task in a fully-volumetric fashion could be the scarcity of training data - for in such a case each volume does not yield many training samples, as for 2D and 3D-patch-based solutions, but a single one. To investigate this possible drawback, we evaluate the performance of CEREBRUM when trained on smaller sub-sets of our database. In particular, we train the proposed model by randomly extracting 25, 50, 100, 250, 500, 700, 900 samples from the training set. To evaluate the performance of the model in the first two cases (i.e., 25 and 50 MRI scans), we repeat the training 5 times (on randomly extracted yet non-overlapping subsets of the database) and average the results. Furthermore, we evaluate the impact on the performance yielded by the introduction of strided convolutions (i.e., more learnable parameters) when the training set size is limited by training a variation of CEREBRUM where max-pooling is used as a dimensionality-reduction strategy. Fig. 7 shows that the performance variation significantly deteriorates as the training set size falls below 250 samples, while substantial stability is reached over 750 samples. This confirms that our 900 samples training set is properly sized for the task, without there being any urge for data augmentation.



(a) Cerebellum (sagittal view). (b) Basal ganglia (longitudinal view).

Fig. 5. Soft segmentation maps of test subject 1 cerebellum (a) and the basal ganglia (b) produced by the best 2D-patch-based model (QuickNAT), the 3D-patch-based model (3D U-Net), and CEREBRUM (ours). The proposed approach produces results that are spatially more coherent, and lack of false positives (highlighted in light blue). Base-10 logarithmic scale of percent probability is used for visualisation purposes (best viewed in electronic format).



(a) Normalised average voxel-wise entropy $^{H_V}/_{H_V^{MAX}}$.

(b) Normalised average voxel-wise cross-entropy $^{CH_V}/_{CH_V^{MAX}}$.

Fig. 6. Normalised average voxel-wise entropy H_V/H_V^{MAX} (a), and normalised average voxel-wise cross-entropy CH_V/CH_V^{MAX} (b). The probability maps resulting from 2D-patch-based (red, green, blue, and grey for longitudinal, sagittal, coronal, and view-aggregation, respectively), the 3D-patch-based (pink), and our models (yellow for max-pooling and orange for strided convolutions) are compared. The height of the bar indicates the mean across all the test subjects, while the error bar represents the standard deviation. The asterisks below the bars highlight statistically significant results ($p < 0.05$), where the p-value is obtained from a paired t-test computed with respect to the strided-convolutions version of CEREBRUM, labelled with "ref." (best viewed in electronic format).

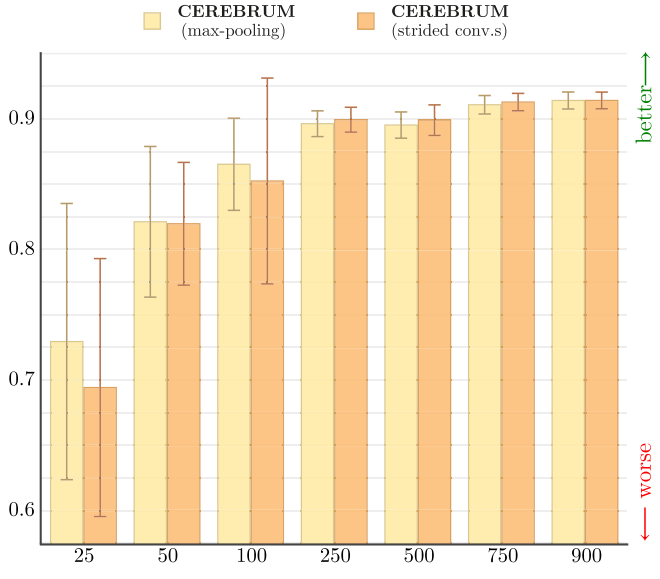


Fig. 7. Impact of the training set size on the performance - Dice Coefficient averaged across all the seven classes. Results are computed on the whole test set (36 volumes).

5. Conclusion

In this work we presented CEREBRUM, a CNN-based deep model that approaches the brain MRI segmentation problem in a fully-volumetric fashion. The proposed architecture is a carefully (architecturally) optimised encoder-decoder that, starting from a $T1_W$ MRI volume, produces a result in only few seconds on a desktop GPU. We evaluated the proposed model performance, comparing it to state-of-the-art 2D and 3D-patch-based models with similar structure, exploiting the Dice Coefficient, the 95th percentile Hausdorff Distance, and the Volumetric Similarity, assessing CEREBRUM superior performance. Furthermore, we conducted a survey of expert neuroscientists to obtain their judgements about the accuracy of the resulting segmentation, comparing the latter with the result of FreeSurfer cortical reconstruction process. According to the participants to such experiment, CEREBRUM achieves better segmentation than FreeSurfer. To our knowledge, this is the first time a DL-based fully-volumetric approach for brain MRI segmentation is deployed. The results we obtained prove the potential of this approach, as CEREBRUM outperforms 2D and 3D-patch-based encoder-decoder models using far less parameters. Removing the partitioning of the volume, as hypothesised, allows the model to learn both local and spatial features. Furthermore, we are also the first conducting a qualitative assessment test consulting expert neuroscientists: this is fundamental, as commonly used metrics often fail to capture the information experts need to rely on DL methods and exploit the latter for research.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Dennis Bontempi: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Sergio Benini:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Project administration. **Alberto Signoroni:** Writing - review & editing, Supervision, Funding acquisition. **Michele Svanera:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Supervision, Project administration. **Lars Muckli:** Resources, Funding acquisition.

Acknowledgments

This project has received funding from the European Unions Horizon 2020 Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2) awarded to LM.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101688.

References

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain MRI segmentation: state of the art and future directions. J. Digit Imaging 30 (4), 449–459. doi:10.1007/s10278-017-9983-4.

Andermatt, S., Pezold, S., Cattin, P., 2016. Multi-dimensional gated recurrent units for the segmentation of biomedical 3d-data. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications. Springer International Publishing, Cham, pp. 142–151. doi:10.1007/978-3-319-46976-8_15.

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D.S., Craddock, R.C., 2017. The neuro bureau ADHD-200 preprocessed repository. Neuroimage 144, 275–286. doi:10.1016/j.neuroimage.2016.06.034.

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., Bach Cuadra, M., 2011. A review of atlas-based segmentation for magnetic resonance brain images. Comput. Methods Programs Biomed. 104 (3), e158–e177. doi:10.1016/j.cmpb.2011.07.015.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.-A., 2018. Voxresnet: deep voxelwise residual networks for brain segmentation from 3d mr images. Neuroimage 170, 446–455. doi:10.1016/j.neuroimage.2017.04.041.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016. Springer International Publishing, Cham, pp. 424–432.

Crdenes, R., de Luis-Garca, R., Bach-Cuadra, M., 2009. A multidimensional segmentation evaluation for medical image data. Comput. Methods Programs Biomed. 96 (2), 108–124. doi:10.1016/j.cmpb.2009.04.009.

Deniz, C.M., Xiang, S., Hallyburton, R.S., Welbeck, A., Babb, J.S., Honig, S., Cho, K., Chang, G., 2018. Segmentation of the proximal femur from mr images using deep convolutional neural networks. Sci. Rep. 8 (1), 16485. doi:10.1038/s41598-018-34817-6.

Despotović, I., Goossens, B., Philips, W., 2015. MRI Segmentation of the human brain: challenges, methods, and applications. Comput. Math. Methods Med. 2015, 450341(1–23). doi:10.1155/2015/450341.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I., 2019. HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imaging 38 (5), 1116–1126. doi:10.1109/TMI.2018.2878669.

Ellingsen, L.M., Roy, S., Carass, A., Blitz, A.M., Pham, D.L., Prince, J.L., 2016. Segmentation and labeling of the ventricular system in normal pressure hydrocephalus using patch-based tissue classification and multi-atlas labeling. In: Styner, M.A., Angelini, E.D. (Eds.), Medical Imaging 2016: Image Processing. SPIE, pp. 116–122. doi:10.1117/12.2216511. International Society for Optics and Photonics.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S.S., Wright, J., Durnez, J., Poldrack, R.A., Gorgolewski, K.J., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16 (1), 111–116. doi:10.1038/s41592-018-0235-4.

Fedorov, A., Johnson, J., Damaraju, E., Ozerin, A., Calhoun, V., Plis, S., 2017. End-to-end learning of brain tissue segmentation from imperfect labeling. In: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 3785–3792. doi:10.1109/IJCNN.2017.7966333.

Fischl, B., 2012. FreeSurfer. Neuroimage 62 (2), 774–781. doi:10.1016/j.neuroimage.2012.01.021.

FreeSurfer, 2008. Recon-all run times. https://surfer.nmr.mgh.harvard.edu/fswiki/ReconAllRunTimes. [Online; accessed 11-September-2019].

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R., 2018. Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. 47, 45–67. doi:10.1016/j.media.2018.03.006.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15 (9), 850–863. doi:10.1109/34.232073.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Klein, A., Ghosh, S.S., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Neto, E.C., Keshavan, A., 2017. Mindboggling morphometry of human brains. PLoS Comput. Biol. 13 (2), e1005350. doi:10.1371/journal.pcbi.1005350.

Lerch, J.P., van der Kouwe, A.J.W., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K.L., Smith, S.M., Fischl, B., Sotiropoulos, S.N., 2017. Studying neuroanatomy using MRI. Nat. Neurosci. 20 (3), 314–326. doi:10.1038/nn.4501.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3d convolutional networks: Brain parcellation as a pretext task. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 348–360.

Li, X., Morgan, P.S., Ashburner, J., Smith, J., Rorden, C., 2016. The first step for neuroimaging data analysis: DICOM to NIfti conversion. J. Neurosci. Methods 264, 47–56. doi:10.1016/j.jneumeth.2016.03.001.

Li, Y., Guo, L., Zhou, Z., 2019. Towards safe weakly supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 1. Early Access.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Snchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. doi:10.1016/j.media.2017.07.005.

Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19 (9), 1498–1507. doi:10.1162/jocn.2007.19.9.1498.

McClure, P., Rho, N., Lee, J. A., Kaczmarzyk, J. R., Zheng, C., Ghosh, S. S., Nielson, D., Thomas, A., Bandettini, P., Pereira, F., 2018. Knowing What You Know in Brain Segmentation Using Deep Neural Networks. arXiv:1812.01719 [cs, stat].

Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., de Bresser, J., Alansary, A., de Bruijne, M., Carass, A., El-Baz, A., Jog, A., Katyal, R., Khan, A.R., van der Lijn, F., Mahmood, Q., Mukherjee, R., van Opbroek, A., Paneri, S., Pereira, S., Persson, M., Rajchl, M., Sarikaya, D., Smedby, O., Silva, C.A., Vrooman, H.A., Vyas, S., Wang, C., Zhao, L., Biessels, G.J., Viergever, M.A., 2015. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput. Intell. Neurosci. 2015, 1–16. doi:10.1155/2015/813696.

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nat. Neurosci. 19. doi:10.1038/nn.4393. 1523 EP –

Milletari, F., Navab, N., Ahmadi, S., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. doi:10.1109/3DV.2016.79.

Oxtoby, N. P., Ferreira, F. S., Mihalik, A., Wu, T., Brudfors, M., Lin, H., Rau, A., Blumberg, S. B., Robu, M., Zor, C., et al., 2019. ABCD Neurocognitive Prediction Challenge 2019: Predicting Individual Residual Fluid Intelligence Scores from Cortical Grey Matter Morphology. arXiv:1905.10834.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56 (3), 907–922. doi:10.1016/j.neuroimage.2011.02.046.

Pawlowski, N., Ktena, S.I., Lee, M.C.H., Kainz, B., Rueckert, D., Glocker, B., Rajchl, M., 2017. DLTK: state of the art reference implementations for deep learning on medical images. ariXiv:1711.06853.

Peirce, J.W., 2007. Psychopy psychophysics software in python. J. Neurosci. Methods 162 (1–2), 8–13.

Quan, T. M., Hildebrand, D. G., Jeong, W.-K., 2016. FusionNet: a deep fully residual convolutional neural network for image segmentation in connectomics. arXiv:1612.05360.

Rajchl, M., Pawlowski, N., Rueckert, D., Matthews, P. M., Glocker, B., 2018. NeuroNet: Fast and Robust Reproduction of Multiple Brain Image Segmentation Pipelines. arXiv:1806.04224.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., ADNI, 2019. QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. Neuroimage 186, 713–727.

Savioli, N., Montana, G., Lamata, P., 2019. V-FCNN: Volumetric fully convolution neural network for automatic atrial segmentation. In: Pop, M., Sermesant, M., Zhao, J., Li, S., McLeod, K., Young, A., Rhode, K., Mansi, T. (Eds.), Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. Springer International Publishing, Cham, pp. 273–281.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15 (1), 29.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., 2013. The WU-minn human connectome project: an overview. Neuroimage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep learning for computer vision: a brief review. Comput. Intell. Neurosci. 2018, 1–13. doi:10.1155/2018/7068349.

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: Deep Convolutional Neural Network for segmenting neuroanatomy. Neuroimage 170, 434–445. doi:10.1016/j.neuroimage.2017.02.035.

Weier, K., Beck, A., Magon, S., Amann, M., Naegelin, Y., Penner, I.K., Thürling, M., Aurich, V., Derfuss, T., Radue, E.-W., Stippich, C., Kappos, L., Timmann, D., Sprenger, T., 2012. Evaluation of a new approach for semi-automatic segmentation of the cerebellum in patients with multiple sclerosis. J. Neurol. 259 (12), 2673–2680. doi:10.1007/s00415-012-6569-4.

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N.C., Kühn, S., Schaefer, S., Heinze, H.-J., Düzel, E., Bäckman, L., Lindenberger, U., Lövdén, M., 2014. Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains: comparing manual and automatic segmentation of hc volumes. Hum. Brain Mapp. 35 (8), 4236–4248. doi:10.1002/hbm.22473.

Zhan, M., Goebel, R., de Gelder, B., 2018. Ventral and dorsal pathways relate differently to visual awareness of body postures under continuous flash suppression. eNeuro 5 (1). doi:10.1523/ENEURO.0285-17.2017.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57. doi:10.1109/42.906424.

Zhou, Z.-H., 2017. A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5 (1), 44–53.